DOCUMENT RESUME

ED 090 288 TM 003 567

AUTHOR Hummel, Thomas J.; Feltovich, Paul J.
TITLE Empirical Sampling Distributions of the Product Moment Correlation Coefficient When Bivariate Observations are Correlated.
PUB DATE 74
NOTE 19p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, Illinois, April 15-19, 1974)

EDRS PRICE MF-$0.75 HC-$1.50 PLUS POSTAGE
DESCRIPTORS Computer Programs; *Correlation; Hypothesis Testing; Matrices; Sampling; Statistical Analysis; *Statistical Bias; *Statistics
IDENTIFIERS Monte Carlo Method

ABSTRACT

In some correlational studies it is not reasonable to assume that bivariate observations are uncorrelated. An example would be a configural analysis in which two individuals are correlated across several variables (e.g., Q-technique). The present study was a Monte Carlo investigation of the robustness of techniques used in judging the magnitude of a sample correlation coefficient when observations are correlated. Empirical distributions of r, t, and Fisher's z were generated. Patterns of correlation were found which caused error rates to be as high as 0.20 when the nominal alpha was 0.05. A technique for controlling error rates in certain situations is suggested. (Author)

ED 090288

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

25.07

EMPIRICAL SAMPLING DISTRIBUTIONS OF THE PRODUCT MOMENT CORRELATION COEFFICIENT WHEN BIVARIATE OBSERVATIONS ARE CORRELATED

by

Thomas J. Hummel and Paul J. Feltovich

University of Minnesota

TM 003 567

A Paper Presented at the Annual Convention
of the
American Educational Research Association
Chicago
1974

SESSION NO. 25.07

ABSTRACT

In some correlational studies it is not reasonable to assume that bivariate observations are uncorrelated. An example would be a configural analysis in which two individuals are correlated across several variables (e.g. Q-technique). The present study was a Monte Carlo investigation of the robustness of techniques used in judging the magnitude of a sample correlation coefficient when observations are correlated. Empirical distributions of r, t, and Fisher's z were generated. Patterns of correlation were found which caused error rates to be as high as .20 when the nominal alpha was .05. A technique for controlling error rates in certain situations is suggested.

Requests for reprints of this article may be sent to Dr. Thomas J. Hummel, Associate Professor and Research Psychologist, Education Career Development Office, College of Education, University of Minnesota, 1425 University Avenue Southeast, Minneapolis, Minnesota 55414.

Introduction

In order to use the distribution of the sample correlation coefficient, r, in testing its statistical significance, one of the necessary assumptions is that the bivariate observations be independent. Situations arise in which this assumption may not be warranted. In taxonomic problems, a Q-correlation is one index which can be used to judge profile similarity. Here people rather than variables are correlated, and the bivariate observations used in the calculation of r are not, in general, independent since each person has a score in each bivariate observation. Another example arises in correlating the observations of two judges who have rated the same person or group on a number of dimensions. An investigator attempting to judge the magnitude of a correlation coefficient in such situations might be unwilling to refer r to Fisher's distributions for they are based on a different model.

Purpose

Consider the n x 2 data matrix $\underset{\sim}{Y}=[y_{ij}]$ in which the rows are indexed by i=1,..., n, and the columns are subscripted by j=1,2. If the rows are randomly drawn observation vectors from a bivariate normal distribution, then the distribution of the sample product moment correlation coefficient between the columns of $\underset{\sim}{Y}$, r_{12}, is known. Fisher (1914) obtained the distributions for both the $\rho_{12}=0$ and $\rho_{12}\neq 0$ cases. Morrison (1962) and MacGregor (1962) studied the distribution of r_{12} when the rows of $\underset{\sim}{Y}$ are not independent. Both authors made restrictive assumptions about the pattern of correlation

coefficients among observations, or values of $\rho_{ij,i'j'}$. The objective of the present study was twofold: 1) to construct a computer program which could generate sample values of r_{12} based on observations from populations in which the population correlation structure describing a n x 2 matrix $\underline{Y}$ could be specified by the user; and 2) to obtain empirical distributions of r_{12} so that the robustness of the techniques used to judge the magnitude of r_{12} could be observed. Since the number of parameters which must be specified in this type of investigation is large (in the n=10 case, for example, 190 parameters must be specified), Monte Carlo methods and limited computing time prohibit the complete specification of a family of distributions. The purpose, therefore, was to determine if error rates could be affected by dependencies among the data and to investigate variables which might relate to any existing lack of robustness.

For convenience, this investigation is discussed using terminology associated with hypothesis testing, such as α-level and error rate. However, there is a direct application of the findings to areas in which hypothesis testing in the usual sense is not of interest. For example, in studies where Q-correlations are based on two randomly drawn observation vectors, the investigator knows that $E(r_{12})=0$; therefore, he is not interested in testing $H_0:\rho_{12}=0$. He still, however, may be interested in determining how extreme an obtained correlation is in the sampling distribution of r_{12}.

Methods

The present Monte Carlo investigation used a Fortran program written for the University of Minnesota's CDC 6600. The program provides distributions of r_{12}, Student's t, and Fisher's z, computes the first four moments, gives plots, and tallies the extreme values of each distribution. Restrictions in the present program are $n \leq 20$ and a given population correlation matrix must be positive definite.

The program user initializes the values of the 2n x 2n matrix $\underline{R}=[\rho_{ij,i'j'}]$, the underlying population correlation matrix for the elements of $\underline{Y}$. The n x n submatrices of $\underline{R}$, $\underline{R}_{11}$ and $\underline{R}_{22}$, specify the population correlations for the elements within the first and second columns of $\underline{Y}$, respectively. $\underline{R}_{12}$ and $\underline{R}_{21}$ specify the population values for relationships between elements in the two columns. The program simulates the situation in which n-dimensional vectors from a multivariate normal population are selected and then correlated.

This investigation builds on a result obtained by Morrison (1962), which using the present notation can be written as

$$\rho^{*}_{12} = \frac{\rho_{i1,i2}-\rho_{i1,i'2}}{[(1-\rho_{i1,i'1})(1-\rho_{i2,i'2})]^{1/2}} .$$

In this expression, ρ^{*}_{12} is a non-centrality parameter for Fisher's distribution of r_{12}, $i \neq i'$, and $\rho_{ij,i'j'}$ values in the equation are constants as i or i' vary from 1 to n. In other words, the distribution

of r_{12} based on correlated observations is known if the off-diagonal elements of $\underline{R}_{11}$, $\underline{R}_{22}$, and $\underline{R}_{12}$ are equal to constants (possibly the same constant) and the diagonal terms of $\underline{R}_{12}$ equal a constant.

In some psychological applications the assumption of constant correlation required in Morrison's equation would be overly restrictive. Therefore, the strategy was to vary the off-diagonals in the submatrices of $\underline{R}$ and to observe the effect on error rates.

Data and Program Verification

The computer program uses a random number generator to create 2n-dimensional vectors of independent standard normal deviates. The elements in these vectors are then randomly permuted to insure adequate coverage of 2n-dimensional space. Each vector is transformed by a triangular factorization of the matrix $\underline{R}$ to obtain a vector $\underline{z} \sim N(\underline{0}, \underline{R})$ (Scheuer & Stoller, 1962) and then partitioned into two n-dimensional vectors, i.e. the two columns of $\underline{Y}$.

For all of the distributions of r_{12} generated: a) $\underline{R}$ was initialized by specifying the values of $\underline{R}_{11}$, $\underline{R}_{22}$, $\underline{R}_{12}$ (see for example Table 5); b) $\rho_{i1,i2}=\rho_{12}$ for all i; and c) n=10. A separate initializing program was written which accepts specifications for initializing $\underline{R}$ and then, if necessary, employs an iterative numerical procedure which reduces the level or dispersion of the correlations to obtain a positive definite matrix.

The program was tested using n=10 and generating correlations which produced known distributions of r_{12}. In the most stringent test, $\underline{R}$ was initialized so that Morrison's equation [1] would provide the exact non-centrality parameter, ρ_{12}, (i.e. $\underline{R}_{11}$, $\underline{R}_{22}$, $\underline{R}_{12}$ conformed

to Morrison's requirements and the distribution of r_{12} was known). The values of $\underline{R}$ were $\rho_{i1,i2}=.513$, $\rho_{i1,i'2}=.227$, $\rho_{i1,i'1}=\rho_{i2,i'2}=.285$, $i\neq i'$, and the non-centrality parameter, ρ^{*}_{12}, was .4. Using Soper, Young, Cave, Lee, and Pearson's (1916) tables with non-centrality parameter $\rho^{*}_{12}=.4$ and n=10, the theoretical moments of the distribution of r_{12} were compared to the obtained empirical moments (see Table 1).

Results

The study was divided into two segments. In the first segment (Case I), $\underline{R}_{11}=\underline{R}_{22}$ and $\underline{R}_{12}=\underline{R}_{21}=\underline{0}$, and in the second (Case II), $\underline{R}_{11}=\underline{R}_{22}$ and $\underline{R}_{12}=\underline{R}_{21}\neq\underline{0}$. Each experiment reported was based on 10,000 realizations of r_{12}.

Case I

Since Morrison's results showed that the distributions of r_{12} remained unchanged for the case assuming constant off-diagonal elements in $\underline{R}_{11}$ and $\underline{R}_{22}$ and $\underline{R}_{12}=\underline{R}_{21}=\underline{0}$, the first question investigated was whether or not heterogeneity in the off-diagonal elements in $\underline{R}_{11}$ and $\underline{R}_{22}$ could cause the actual alpha level to be substantially different from the nominal value. The correlations presented in Table 2 were used in both $\underline{R}_{11}$ and $\underline{R}_{22}$ in a Monte Carlo run designed to answer this question. Table 3 presents empirical and theoretical moments and gives the empirical alpha level obtained when the critical value was $t_{.05/2}$ with eight degrees of freedom. It is apparent from the moments that while the distribution based on correlated observations remains centered at zero and unskewed, it is more variable and platykurtic causing the obtained α to be more than twice as large as the nominal level.

TABLE 1

PROGRAM TEST. COMPARISON OF THEORETICAL AND OBTAINED MOMENTS AND ERROR RATES FOR THE CASE $\rho_{12}=\rho_{12}^{*}=.4$

	ERROR RATE	MEAN	VARIANCE	β_1	β_2
OBTAINED	.0500	.3795	.0853	.4337	3.1307
THEORETICAL	.0500	.3813	.0851	.4374	3.1669

TABLE 2

CASE 1. CORRELATIONS USED IN RUN IN WHICH $\underline{R}_{11} = \underline{R}_{22}$ HAD HIGHEST HETEROGENEITY

1.000	.725	.725	.725	.725	.725	.725	.725	.275	.275
	1.000	.725	.725	.725	.725	.725	.275	.275	.275
		1.000	.725	.725	.725	.725	.275	.275	.275
			1.000	.725	.725	.725	.275	.275	.275
				1.000	.725	.725	.275	.275	.275
					1.000	.725	.275	.275	.275
						1.000	.275	.275	.275
							1.000	.275	.275
								1.000	.275
									1.000

TABLE 3

CASE 1. COMPARISON OF MOMENTS AND ERROR RATES FOR THE $\rho_{12} = 0$ THEORETICAL SAMPLING DISTRIBUTION AND THE EMPIRICAL SAMPLING DISTRIBUTION OBTAINED WHEN $\underline{R}_{11} = \underline{R}_{22}$ HAD HIGHEST HETEROGENEITY.

	ERROR RATE	MEAN	VARIANCE	β_1	β_2
OBTAINED	.1130	-.0020	.1576	.0002	2.2246
THEORETICAL	.0500	.0000	.1111	.0000	2.4545

The heterogeneity of the correlations in Table 2 approaches a maximum given the constraint mentioned in the data section above. $\underset{\sim}{R}_{11}=\underset{\sim}{R}_{22}$ were hand set with highly heterogeneous values and iterated by the initializing program until the positive definiteness criterion was met. Because it was believed that these extremely variant correlations were unrealistic for many types of psychological data, it was decided to investigate matrices with less variance among the $\rho_{ij,i'j'}$s. Two levels of variance were chosen, $\sigma^2_L=.004$ and $\sigma^2_H=.016$, where σ^2 is based on the off-diagonals of $\underset{\sim}{R}_{11}=\underset{\sim}{R}_{22}$. Mean levels of correlation were studied within each level of variance to determine if given a particular level of heterogeneity among the correlations, the average level of correlation would have an effect ($\mu_L=.285$ and $\mu_H=.569$). Table 4 presents the obtained moments for these four runs and the actual α levels obtained when a critical value of $t_{.05/2}$ with eight degrees of freedom was used.

It is clear from these results that variance plays the predominant role in affecting the error rates and that once a degree of heterogeneity is established, the mean level of correlation also has some effect. Table 5 contains the correlations used to initialize $\underset{\sim}{R}_{11}=\underset{\sim}{R}_{22}$ for low mean and low variance so that the reader might contrast the correlation matrices used for the highest and the lowest obtained error rates in Case I.

Case II

As with Case I, the computer runs in Case II were designed to deviate from Morrison's findings in that the correlation coefficients in $\underset{\sim}{R}_{11}=\underset{\sim}{R}_{22}$ and the off-diagonals of $\underset{\sim}{R}_{12}=\underset{\sim}{R}_{21}$ were not equal to

TABLE 4

CASE 1. COMPARISON OF ERROR RATES AND MOMENTS FOR THE EMPIRICAL SAMPLING DISTRIBUTIONS OF r_{12} OBTAINED WHEN LEVEL, μ, AND HETEROGENEITY, σ^2, WERE VARIED

	ERROR RATES	MOMENTS			
		MEAN	VARIANCE	β_1	β_2
σ^2_L μ_L	.0517	.0036	.1128	.0000	2.4289
σ^2_L μ_H	.0536	.0013	.1147	.0003	2.4485
σ^2_H μ_L	.0567	.0016	.1168	.0002	2.4314
σ^2_H μ_H	.0695	-.0041	.1257	.0000	2.3531

TABLE 5

CASE 1. CORRELATIONS IN $\underline{R}_{11} = \underline{R}_{22}$ FOR RUN WHICH YIELDS THE LOWEST ERROR RATE IN CASE 1.

1.000	.371	.371	.371	.371	.255	.255	.255	.255	.255
	1.000	.371	.371	.371	.255	.255	.255	.255	.255
		1.000	.371	.371	.313	.313	.313	.255	.255
			1.000	.371	.313	.313	.313	.313	.313
				1.000	.313	.313	.313	.313	.313
					1.000	.197	.197	.197	.197
						1.000	.197	.197	.197
							1.000	.197	.197
								1.000	.197
									1.000

constants. Given Morrison's findings and the results from Case I, two hypotheses were employed in designing computer runs for Case II: 1) if an analog to Morrison's non-centrality parameter, say ρ'_{12}, is computed by replacing the constants which he specifies with the averages of the off-diagonal elements in $\underset{\sim}{R}_{11}=\underset{\sim}{R}_{22}$ and $\underset{\sim}{R}_{12}=\underset{\sim}{R}_{21}$, then the discrepancy between ρ'_{12} and the diagonal of $\underset{\sim}{R}_{12}=\underset{\sim}{R}_{21}$, say $\delta=|\rho'_{12}-\rho_{i1,i2}|=|\rho'_{12}-\rho_{12}|$, should relate to variability among the error rates (This hypothesis was motivated by inspection of Morrison's equation which shows that when $\underset{\sim}{R}_{12}\neq\underset{\sim}{0}$ the actual location of the distribution is changed when dependencies across bivariate observations exist.); and 2) given a value δ, variability of the correlations around the average values substituted into Morrison's equation should also explain variability among the error rates.

The conditions studied were $\delta_1=.006$, $\delta_2=.035$, $\delta_3=.158$ and $\delta_4=.262$. Nested within each level of δ were two levels of variance, σ^2_H and σ^2_L. Matrices for the σ^2_L case were created by holding the mean levels of the off-diagonals of $\underset{\sim}{R}_{11}=\underset{\sim}{R}_{22}$ and $\underset{\sim}{R}_{12}=\underset{\sim}{R}_{21}$ constant, so that ρ'_{12} would be unaffected, and then halving the standard deviation of these off-diagonals (see Table 6).

The error rates obtained from these eight runs varied from .0527 to .2011. They were obtained by computing the proportion of correlations which exceeded a critical value based on Fisher's z-transformation, and therefore simulate the situation in which a researcher tests the significance of an obtained correlation at a nominal $\alpha=.05$ using Fisher's model (i.e. where $\underset{\sim}{R}_{11}=\underset{\sim}{R}_{22}=I$ and

TABLE 6

CASE 2. $\rho_{i1,i2}$, MEANS, AND VARIANCES FOR THE MATRICES OF CASE 2.

		δ_1=.006	δ_2=.035	δ_3=.158	δ_4=.262
MEAN	$\underline{R}_{11}$	.5693	.2843	.2854	.5694
	$\underline{R}_{12}$	.4528	.1678	.2272	.5112
σ_L^2	$\underline{R}_{11}$	.0039	.0039	.0010	.0010
	$\underline{R}_{12}$	.0039	.0039	.0010	.0010
σ_H^2	$\underline{R}_{11}$	.0157	.0157	.0039	.0039
	$\underline{R}_{12}$	.0156	.0156	.0039	.0039
$\rho_{12} = \rho_{i1,i2}$		.800	.500	.400	.700

$\underset{\sim}{R}_{12}=\underset{\sim}{R}_{21}$ is diagonal with nonzero elements equal to ρ_{12}). Critical values for the z-statistic were based on moments given by Kendall and Stuart (1963) rather than the approximations typically found in applied statistics texts. The results in Table 7 clearly show that a researcher may be working at an α level much higher than the announced value. The discrepancy, $\delta=|\rho'_{12}-\rho_{12}|$, accounts for most of the discrepancy among the error rates. Variance has little effect relative to this discrepancy. In fact, if ρ'_{12} had been used in obtaining cutoff points rather than ρ_{12}, the rejection rates for the four Case II matrices producing the highest error rates, .201, .198, .084, and .083, would have been .057, .053, .055, and .052, respectively.

Conclusions

When a researcher has correlation coefficients based on data in which a degree of correlation exists not only between the columns of the data matrix but also among the rows and he wishes to judge the magnitude of such correlations, he should not assume that values based on the t-distribution (if the hypothesis is $\rho_{12}=0$) or on Fisher's z-transformation (if the hypothesis is ρ_{12}=some constant) will give him a test of size α. There are situations in which the error rate can be four times that of the nominal α and, of course, there may be situations other than those investigated in which the situation is actually worse.

What has been learned is that variability among the correlations in the off-diagonals of $\underset{\sim}{R}_{11}=\underset{\sim}{R}_{22}$, given $\underset{\sim}{R}_{12}=\underset{\sim}{R}_{21}=\underset{\sim}{Q}$, can affect the error rates and, given that variability exists, the

TABLE 7

CASE 2. ERROR RATES AND MOMENTS FOR THE EMPIRICAL SAMPLING DISTRIBUTIONS OF r_{12}.

		ERROR RATES	MOMENTS			
			MEAN	VARIANCE	β_1	β_2
δ_1	σ_L^2	.0527	.7851	.0210	3.1988	8.3180
	σ_H^2	.0619	.7789	.0237	3.0181	7.4546
δ_2	σ_L^2	.0540	.4421	.0777	.6267	3.5143
	σ_H^2	.0600	.4365	.0812	.5914	3.3506
δ_3	σ_L^2	.0833	.2280	.1037	.1570	2.6611
	σ_H^2	.0840	.2222	.1032	.1360	2.6650
δ_4	σ_L^2	.1977	.4193	.0802	.5168	3.2785
	σ_H^2	.2011	.4174	.0834	.5305	3.2746

level of correlation also plays a small role. In the non-null case, $\underline{R}_{12}=\underline{R}_{21}\neq\underline{0}$, it was found that heterogeneous correlations in the off-diagonals of $\underline{R}_{11}=\underline{R}_{22}$ and $\underline{R}_{12}=\underline{R}_{21}$ affect error rates. However, there is utility in using cutoff points based on ρ'_{12} rather than ρ_{12}. These cutoff points resulted in error rates close to α.

Of course, the results of this study should only be applied to situations in which researchers have correlations similar to those investigated here. It is therefore recommended that researchers use the Monte Carlo method to study the effect of the correlation patterns which underlie the data they tend to investigate. A copy of the programs used in the present investigation may be obtained from the authors.

References

Fisher, R. A. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. Biometrika, 1914, 10, 507-521.

Kendall, M. G. & Stuart, A. The advanced theory of statistics, Vol. 1, Distribution theory (2nd ed.), London: Charles Griffin & Co., 1963.

McGregor, J. R. The approximate distribution of the correlation between two stationary linear Markov series. Biometrika, 1962, 49, 379-388.

Morrison, Donald F. On the distribution of the sums of squares and cross products of normal variates in the presence of intraclass correlation. Annals of Mathematical Statistics, 1962, 33, 1461-1463.

Scheuer, E. M. & Stoller, D. S. On the generation of normal random vectors. Technometrics, 1962, 4, 278-281.

Soper, H. S., Young, A. W., Cave, B. M., Lee, A., & Pearson, K. On the distribution of the correlation coefficient in small samples. Appendix II to the papers of "Student" and R. A. Fisher. A cooperative study. Biometrika, 1916, 11, 328-413.